Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-022-03892-z

ORIGINAL RESEARCH

Traffic condition estimation and data quality assessment for signalized road networks using massive vehicle trajectories

Peixiao Wang[1] · Tong Zhang[1] · Tao Hu[2,3]

Received: 23 July 2021 / Accepted: 28 April 2022

Abstract
Monitoring traffic conditions on urban signalized road networks is an essential component of urban traffic control systems. Due to the sparseness of trajectory data and the influence of signal timing, it is challenging to estimate the traffic condition of large-scale urban signalized networks based on trajectory data. In this study, a novel and integrated data-driven learning approach (NEI-SE) is proposed, incorporating road network segmentation, speed matching, and sparse data imputation for the estimation of travel speed. First, the urban traffic network is divided according to signalized intersection and road segment length, considering the influence of signal timing on urban traffic speed. Then, based on taxi trajectory data and the divided road network, a traffic condition matrix is constructed describing the road conditions. Finally, a lightweight multi-view learning method that integrates temporal patterns and spatial topological relations is proposed to fill the missing values of the traffic condition matrix. The approach was validated on real-world traffic trajectory data collected in Wuhan, China. The results showed that NEI-SE outperformed nine existing baselines in terms of imputation accuracy. In addition, the AutoNavi congestion data was used to evaluate the data quality of the estimated traffic speed data due to lack ground truth of traffic speed. The results showed that the congestion index data had a significant negative correlation with imputed traffic speed series, with an average correlation coefficient of −0.67, proving that the traffic speed data estimated by the proposed approach have satisfactory quality.

Keywords Signalized road networks · Traffic condition estimation · Spatiotemporal data imputation · Traffic flow missing

1 Introduction

Monitoring traffic conditions on urban signalized road networks is an essential component of urban traffic control systems (Angayarkanni et al. 2021; Guo et al. 2019a, b; Li et al. 2020; Younes 2021). In recent years, many fixed-location sensors, such as microwave sensors and loop detectors, have been installed to monitor traffic conditions continuously and collaboratively (Ma et al. 2017; Vigos and Papageorgiou 2010). However, due to the high cost of sensor installation and maintenance, fixed location sensors are unsuitable for large-scale monitoring of urban traffic conditions, especially for complex signalized road networks (González et al. 2020; Hara et al. 2018; Zhang et al. 2020). Fortunately, with the rapid development of connected vehicle technologies and the emergence of agent driving services, a large amount of vehicle trajectory data can be collected, providing a new alternative for traffic condition estimation of urban signalized road networks (Guo et al. 2019a, b; Wang et al. 2022a, b; Yu et al. 2020).

At present, estimation of urban traffic conditions based on trajectory data has been widely investigated in urban planning (Xie et al. 2020), traffic management (Praveen and Raj 2021; Yu et al. 2018), environmental monitoring (Cheng et al. 2020a; b, c, 2021), and other fields. In this study, our goal is to estimate the traffic speed of large-scale urban

✉ Tong Zhang
zhangt@whu.edu.cn

Peixiao Wang
peixiaowang@whu.edu.cn

Tao Hu
taohu.gis@gmail.com

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

[2] Department of Geography, Oklahoma State University, Stillwater, OK 74078, USA

[3] Center for Geographic Analysis, Harvard University, Cambridge, MA 02138, USA

Published online: 03 June 2022

signalized networks, which is a challenging task despite that enormous efforts have been made on this topic (Tao et al. 2012; Yu et al. 2020). The main challenges are as follows.

(1) Complex traffic conditions of signalized road network (Tang et al. 2020a, b): Compared with highway road networks, urban signalized networks have more complex topologies and spatial layouts in signalized intersections. As shown in Fig. 1, due to the influence of signal control, the traffic patterns of road segments l_1 and l_2 are significantly different. In addition, the speeds of road segments l_3 and l_4 are different even if they do not pass through the signalized intersection 5.
(2) Sparse distribution of trajectory data (Wang et al. 2014; Wilby et al. 2014; Zhao et al. 2019): In a specific time window (e.g., 5, 10, or 15 min), the coverage of trajectory data on the entire traffic network is limited, making it difficult to estimate the traffic speed of some traffic segments where trajectory data are sparse or totally missing. For example, at most 30–40% of all road links are covered by trajectory data within 5 min in central Bangkok (Hara et al. 2018).
(3) Quality evaluation of estimation data (Zhan et al. 2017): The lack of ground truth (i.e., the "true" speed of city roads) makes it challenging to evaluate the quality of estimated traffic speed for large-scale urban signalized road networks.

This study addresses the above issues by developing a novel, efficient, and integrated data-driven learning approach (NEI-SE), which incorporates road network segmentation, speed matching, and sparse data imputation to estimate travel speed using vehicle trajectory data. This study makes the following three contributions.

(1) A road network segmentation algorithm called ISD-RoadSeg was proposed to divide the urban road network according to signalized intersection locations and road segment length. In addition, we re-constructed the road network topology after road network segmentation to facilitate speed estimation.
(2) A novel and hybrid learning approach was proposed to estimate the traffic speed of urban signalized road networks. The proposed approach not only considers the spatial topological relationship of the urban road network, but also considers the prior knowledge of the temporal periodicity and temporal proximity of the traffic flow.
(3) Multi-source data, including the AutoNavi congestion data, were used to evaluate the quality of the estimated traffic speed. In addition, we release the estimated traffic speed dataset to support related research on traffic condition imputation and prediction on the urban signalized road network.

2 Related works

In this section, we first review the research related to traffic condition estimation based on trajectory data, and then review sparse traffic data imputation methods.

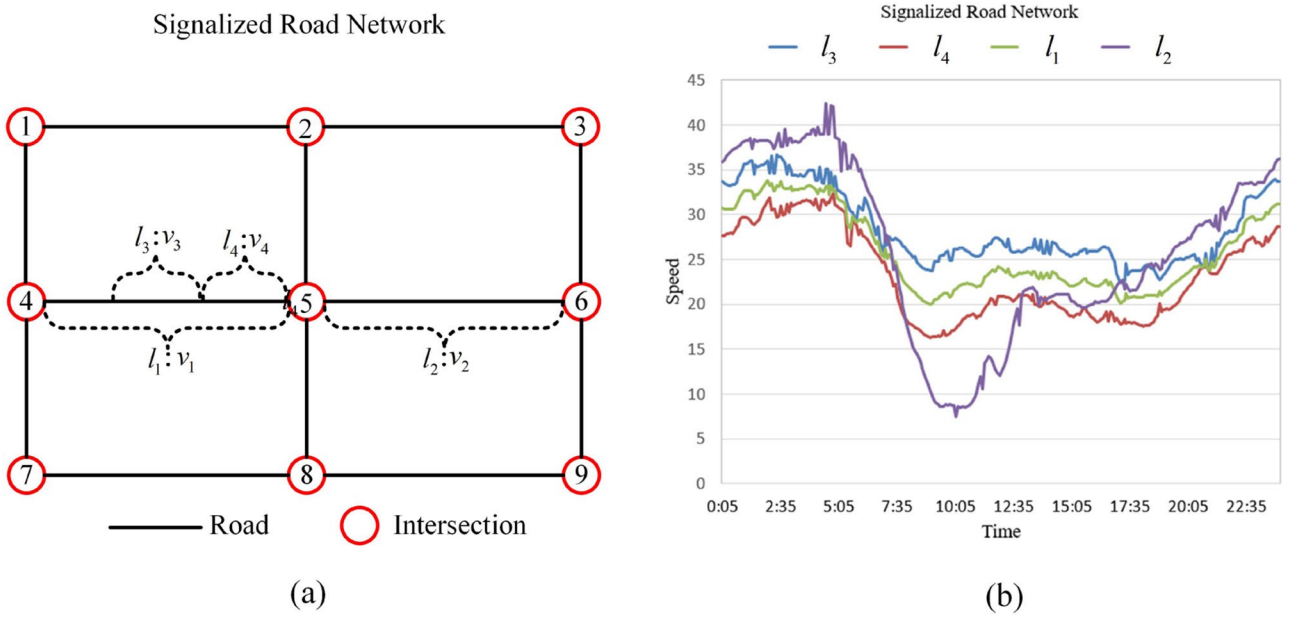

Fig. 1 The challenge of speed estimation on signalized road network

2.1 Traffic condition estimation based on trajectory data

In past years, several studies have been made to estimate urban traffic condition using sampled vehicle trajectories. For example, Tao et al. (2012) presented a microscopic traffic condition estimation method, which collected real-time position data through assisted global positioning system (A-GPS) mobile phones to estimate traffic speed on urban roads. Zhang et al. (2020) proposed a novel traffic flow estimation model called TGMC-S by combining camera detection data and floating vehicle data. They defined a spatial smoothing index (SSI) to measure the difficulty of traffic flow estimation on each segment. Hara et al. (2018) proposed a mixture gaussian graphical model (Mixed-GGM) model based on probe trajectory data. Zhan et al. (2017) constructed a flow-speed diagram (Q–V diagram) based on taxi trajectory and video data, and established travel speed estimation (TSE) and traffic volume inference (TVI) models. In addition to estimating urban traffic speed and volume, some scholars also used trajectory data to estimate traffic pollution emission (Shang et al. 2014) and vehicle queue length (Zhan et al. 2015; Zhao et al. 2019). However, most of the above studies focused on urban highway networks rather than urban signalized road networks. Tang et al. (2020a, b) used probe trajectory data and signal cycle data to estimate the total traffic volume within a certain period at the intersection level on signalized road networks. However, due to the difficulty of obtaining signal cycle data, this method is still unsuitable for estimating the traffic condition on large-scale road networks.

2.2 Sparse traffic data imputation methods

Sparse traffic data imputation methods can be roughly divided into two categories: statistical learning methods and data-driven methods. Statistical learning methods assume that missing data obey specific mathematical rules in space and time dimensions and establish specific paramatric models to describe the patterns of missing data, such as, inverse distance interpolation (IDW), spatiotemporal inverse distance interpolation (ST-IDW) (Li et al. 2014), spatiotemporal kriging (ST-Kriging) (Aryaputera et al. 2015), autoregressive integrated moving average (ARIMA) (Yozgatligil et al. 2013), simple exponential smoothing (SES) (Gardner 2006), and point estimation model of biased sentinel hospitals-based area disease estimation (P-BSHADE) (Hu et al. 2013; Xu et al. 2013). Although classical statistical methods have been widely used in missing traffic data imputation, achieving good results is still difficult. On the one hand, classical statistical methods are based on strict assumptions, which may not hold in actual traffic environments. On the other hand, traffic flow data have complex spatiotemporal patterns, which are difficult to capture using specific and fixed mathematical formulas (Cheng et al. 2019, 2020a, b, c). Data-driven methods do not require to obey specific mathematical rules but establish nonparametric models to automatically mine spatiotemporal characteristics and impute missing values. For example, Yu et al. (2016a, b) integrated time dependence into the traditional matrix factorization model and proposed a new temporal regularized matrix factorization method (TRMF) to estimate missing values. Chen et al. extended matrix factorization to tensor factorization, mining missing patterns in traffic flow data from a higher-dimensional perspective to fill in missing values (Chen et al. 2018, 2020; Chen and Sun 2021). In addition, relevant studies have applied deep learning algorithms to reconstruct missing data and achieved good results. For instance, Cheng et al. (2020a, b, c) used an extreme learning machine (ELM) to integrate IDW and SES algorithms and proposed a lightweight missing data interpolation model. Cheng and Lu (2017), Jiang et al. (2018) combined deep neural networks and P-BSHADE algorithm to propose a hybrid two-step estimation approach. Tang et al. (2020a, b) proposed to integrate the fuzzy rough set theory and fuzzy neural networks to complete missing traffic flow data. Yang et al. (2021) proposed a bidirectional attention model called ST-LBAGAN based on the generative adversarial network. Xu et al. (2020) proposed the GE-GAN model based on graph embedding and generative adversarial network. Compared with classical statistical methods, data-driven methods do not require prior knowledge and explicit mathematical expressions and have more robust imputation results. However, the above studies still mainly focused on highway networks rather than signalized road networks.

3 Preliminaries and problem definitions

Definition 1 (Trajectory) A trajectory $T = \{p_i\}_{i=1}^{n}$ is an ordered sequence of points $p_i = (id, t_i, x_i, y_i)$, where id is a unique vehicle identifier; t_i is the time at which p_i was collected; (x_i, y_i) correspond to the longitude and latitude, respectively, of a sampled trajectory point at time t_i; and n indicates the number of trajectory points included in the trajectory T.

Definition 2 (Traffic network) A traffic network $G =< \boldsymbol{L}, \mathcal{T} >$ is a directed graph, where $\boldsymbol{L} = \{l_i\}_{i=1}^{m}$ represents a collection of road segments, and $\mathcal{T} = \{\sqcup_i\}_{i=1}^{h}$ represents a collection of topological relations. Considering the influence of traffic signal on traffic speed, we divide G according to specific rules. The segmented road network is expressed as $G_s =< \boldsymbol{L}_s = \{l_i^s\}_{i=1}^{M}, \mathcal{T}_s = \{\sqcup_i^s\}_{i=1}^{H} >$, where $M > m, H > h$.

Definition 3 (Traffic condition matrix) The traffic flow data extracted from the trajectories are used to build a spatiotemporal condition matrix with missing data $\boldsymbol{X} \in \mathcal{R}^{2M*N}$, where $\{s_i\}_{i=1}^{2M}$ represents the spatial dimension of data, $\{w_k\}_{k=1}^{N}$ represents the time dimension of data. As shown in Fig. 2, if $\boldsymbol{X}(s_4, w_1) = \phi$, it means that the traffic speed of road segment l_2 in direction 2 at time window w_1 cannot be obtained from trajectories. The missing condition matrix $\boldsymbol{X}$ is imputed to obtain a complete condition matrix $\widehat{\boldsymbol{X}} \in \mathcal{R}^{2M*N}$, where $\widehat{\boldsymbol{X}}(s_i, w_k) \neq \phi | \forall (i,k)$.

Our goal is to obtain a complete condition matrix based on trajectory and road network data, and to evaluate the data quality of the imputed traffic condition matrix, described in Formula (1).

$$\begin{cases} < \widehat{\boldsymbol{X}}, G_s >= M \leftarrow < \{T_i\}, G > \\ \textit{Evaluate data quality of } \widehat{\boldsymbol{X}} \end{cases} \tag{1}$$

where $\{T_i\}$ represents a set of vehicle trajectories; G represents the traffic network before segmentation; $\mathcal{M}$ represents the traffic speed estimation approach proposed; G_s represents the segmented traffic network; $\widehat{\boldsymbol{X}}$ represents the imputed traffic condition matrix.

4 Methodology

Our approach focuses on traffic speed estimation using massive vehicle trajectories, similar to the approach of Yu et al (2020). The different from Yu et al.'s approach is that we focus on the signalized road networks, and evaluate the data quality of the extracted speed. As shown in Fig. 3, the proposed speed estimation approach is mainly divided into three steps. First, the urban traffic network is divided according to the signalized intersection location and road segment distance, considering the influence of signal timing on urban traffic speed. Then, the traffic speed information on the road segment is extracted by considering the moving direction of trajectories. Finally, a lightweight multi-view learning method is proposed to estimate the missing traffic speed by integrating the temporal periodicity and closeness patterns and the spatial topological connection relations.

4.1 Signalized road network segmentation

The traditional traffic network is unsuitable for extracting traffic condition information on road segments. First, the traffic conditions of signalized networks are severely affected by signal timing, and a single road segment in the physical world often spans multiple signalized intersections. Second, when a road segment is particularly the length, the traffic conditions on different parts of the road segment may also be different. Therefore, to accurately extract the traffic

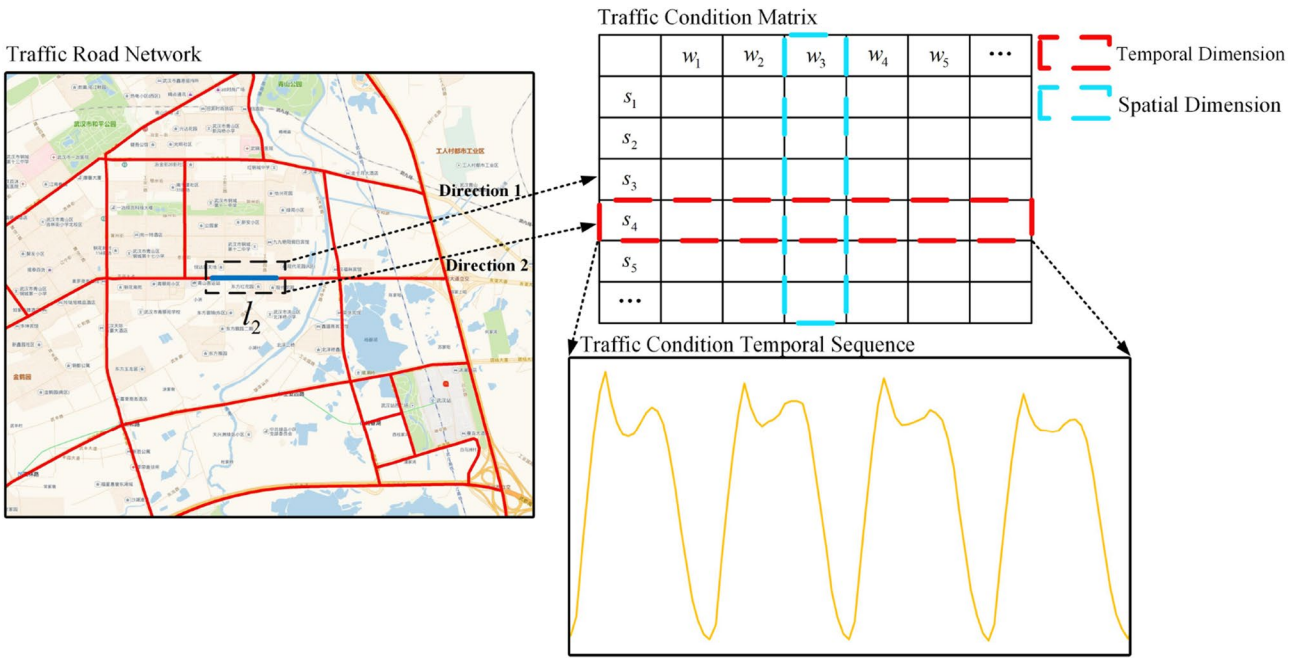

Fig. 2 Problem definitions

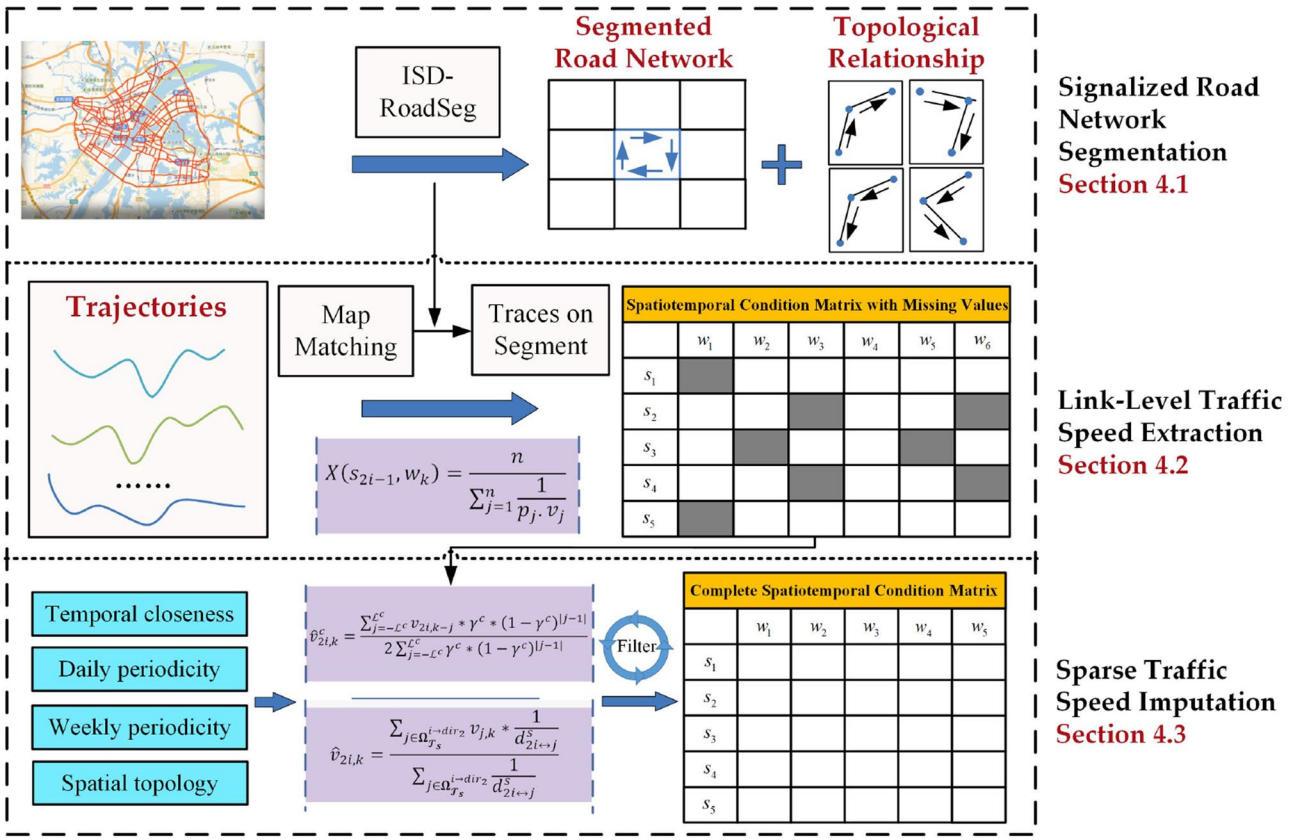

Fig. 3 The approach of trajectories-based traffic speed estimation for signalized road networks

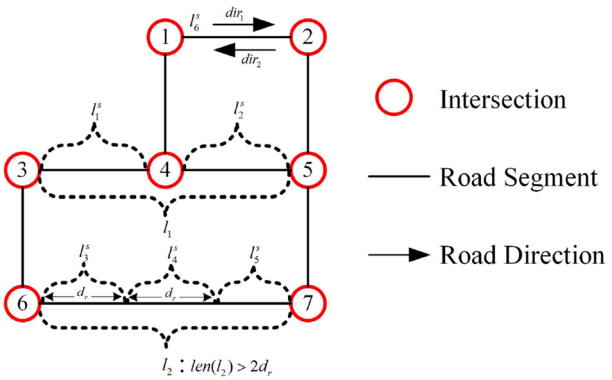

Fig. 4 An illustration of signalized road network segmentation

conditions of the road network, a road network segmentation algorithm considering intersections and segment distance (ISD-RoadSeg) was proposed.

The ISD-RoadSeg algorithm divides the road network $G =< \boldsymbol{L}, \mathcal{T} >$ into $G_s =< \boldsymbol{L_s}, \mathcal{T}_s >$, consisting of two steps: (1) Segmenting road, i.e. $\boldsymbol{L} \rightarrow \boldsymbol{L_s}$, and (2) Reconstructing topological structure $\mathcal{T}_s$

Figure 4 shows the process of road segmentation. First, each road segment in the set $\boldsymbol{L}$ is divided at the signalized intersection. For example, l_1 in the set $\boldsymbol{L}$ is divided into $l_1^s, l_2^s \in \boldsymbol{L_s}$ at the intersection. Then, when the length of the segmented road exceeds $2d_r$, the road segment will be further divided. That is, starting from the start point of the segment, gradually divide the road section of length d_r, and the length of the remaining road segment needs to be greater than d_r and less than $2d_r$. For example, l_2 is divided into l_3^s, l_4^s, and $l_5^s \in \boldsymbol{L_s}$, where the lengths of l_3^s and l_4^s are both d_r. Finally, driving directions are assigned to each resultant road segment. In this study, the direction of a road segment is defined as the clockwise angle between the line connecting the two adjacent points of the road and the geographical north. As each road segment has two directions, we assign the directional information as attributes to each road segment, i.e., dir_1 and dir_2. Taking road segment l_6^s as an example, the calculation methods of $l_6^s \cdot dir_1$ and $l_6^s \cdot dir_2$ are shown in Eqs. (2) and (3).

$$l_6^s \cdot dir_1 = arccos\left(\frac{l_6^s \cdot E \cdot y - l_6^s \cdot S \cdot y}{len(l_6^s)}\right) \tag{2}$$

$$l_6^s \cdot dir_2 = l_6^s \cdot dir_1 + 180° \tag{3}$$

where $\left(l_6^s \cdot S \cdot x, l_6^s \cdot S \cdot y\right)$ represents the start point of road segmemt l_6^s;$\left(l_6^s \cdot E \cdot x, l_6^s \cdot E \cdot y\right)$ represents the end point

of road segmemt l_6^s. For the convenience of calculation, we enforce the constraint that $l_6^s \cdot E \cdot x$ is always greater than $l_6^s \cdot S \cdot x$; *len* is a function used to calculate the length between the start point and the end point of the road segment l_6^s. From the Eqs. (2) and (3), it can be seen that $0° \leq l_6^s \cdot dir_1 < 180°$, and $180° \leq l_6^s \cdot dir_2 < 360°$.

After the road network is divided, the topological structure $\mathcal{T}_s = \left\{\sqcup_i^s\right\}_{i=1}^H$ of $\boldsymbol{L}_s$ is constructed. As shown in Fig. 5, if there is a topological relationship $\sqcup^s =< l_1^s \cdot dir_1 \rightarrow l_2^s \cdot dir_1 >$, it means that the traffic condition on the direction dir_1 of the road segment l_1^s can be transferred to the direction dir_1 of the road segment l_2^s, such as traffic volume. The two adjacent road segments can have at most four types of topological relationships, i.e., $l_1^s \cdot dir_1 \rightarrow l_2^s \cdot dir_1$,$l_1^s \cdot dir_1 \rightarrow l_2^s \cdot dir_2$, $l_1^s \cdot dir_2 \rightarrow l_2^s \cdot dir_2$, and $l_1^s \cdot dir_2 \rightarrow l_2^s \cdot dir_1$. The topological relation construction method of segment l_i^s and segment l_j^s is shown in Eq. (4).

$$\sqcup^s = \begin{cases} l_i^s \cdot dir_1 \rightarrow l_j^s \cdot dir_1 \;\; l_i^s \cdot E = l_j^s \cdot S \\ l_i^s \cdot dir_1 \rightarrow l_j^s \cdot dir_2 \;\; l_i^s \cdot E = l_j^s \cdot E \\ l_i^s \cdot dir_2 \rightarrow l_j^s \cdot dir_2 \;\; l_i^s \cdot S = l_j^s \cdot E \\ l_i^s \cdot dir_2 \rightarrow l_j^s \cdot dir_1 \;\; l_i^s \cdot S = l_j^s \cdot S \end{cases} \tag{4}$$

where $l_i^s \cdot S$ and $l_j^s \cdot S$ represent the start point coordinates of segment l_i^s and l_j^s, respectively; $l_i^s \cdot E$ and $l_j^s \cdot E$ represent the end point coordinates of segment l_i^s and l_j^s, respectively. Algorithm 1 describes the entire procedure of ISD-RoadSeg.

Algorithm 1. ISD-RoadSeg algorithm

Require: Road network: $G =< \boldsymbol{L}, \boldsymbol{\mathcal{T}} >$
Distance threshold: d_r

Ensure: Segmented road network: $G_s =< \boldsymbol{L_s}, \boldsymbol{\mathcal{T}_s} >$

Initialize road segment collection $\boldsymbol{L_s}$ and topological collection $\boldsymbol{\mathcal{T}_s}$
For next $l_i \in \boldsymbol{L}$ **do**
  Divide road segment l_i into $\{sub\_l_i\}$ through the intersections
  while $sub\_l_i \in \{sub\_l_i\}$ **do**
    if $len(sub\_l_i) > 2d_r$ **then**
      Divide road segment sub_l_i into $\{sub\_sub\_l_i\}$ based on d_r
      Add the segmented road segments $\{sub\_sub\_l_i\}$ into $\boldsymbol{L_s}$
    else
      Add the road segment sub_l_i into $\boldsymbol{L_s}$
For next $l_i^s \in \boldsymbol{L_s}$ **do**
  Calculate directions of road segment l_i^s using Equations (2) and (3)
For next $l_i^s \in \boldsymbol{L_s}$ **do**
  For next $l_j^s \in \boldsymbol{L_s}$ **do**
    Construct topological relationship t^s using Equation (4)
    $\boldsymbol{\mathcal{T}_s}$**.add(**$t^s$**)**
return $G_s =< \boldsymbol{L_s}, \boldsymbol{\mathcal{T}_s} >$

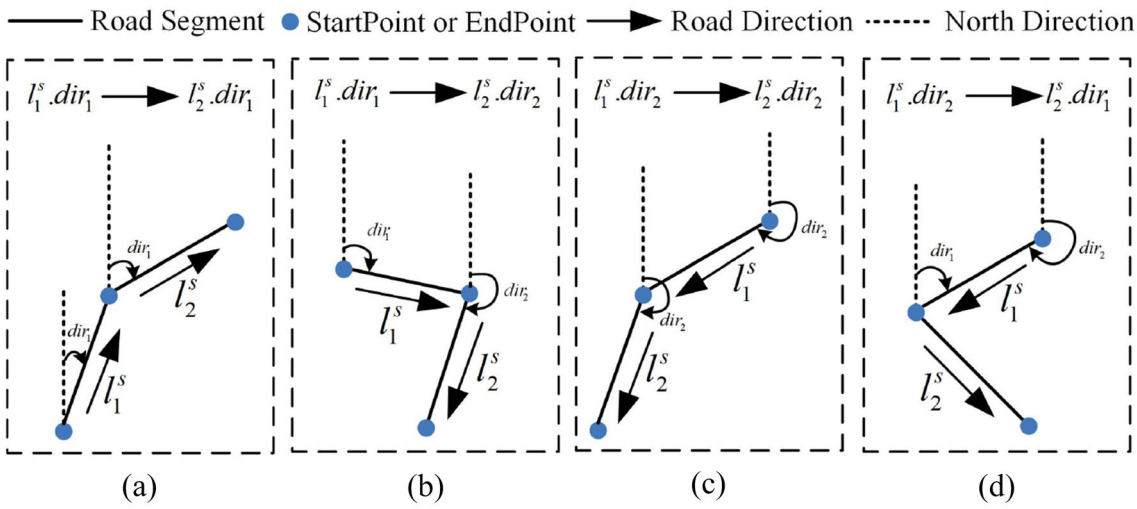

Fig. 5 Four topological relationships of adjacent road segments

4.2 Link-level traffic speed extraction

After the road network is divided, the traffic speed of each road segment is extracted based on the trajectory data, i.e., the traffic condition matrix $\boldsymbol{X}$ is generated. Considering that the collected trajectory points do not contain the vehicle driving direction and speed, the speed and direction of the points are first calculated. The calculation method is shown in Eqs. (5) and (6).

$$p_i \cdot dir_i = \begin{cases} arccos\left(\frac{p_{i+1} \cdot y_{i+1} - p_i \cdot y_i}{len(p_{i+1}, p_i)}\right) & p_{i+1} \cdot x_{i+1} \geq p_i \cdot x_i \\ arccos\left(\frac{p_i \cdot y_i - p_{i+1} \cdot y_{i+1}}{len(p_{i+1}, p_i)}\right) + 180° & p_{i+1} \cdot x_{i+1} \leq p_i \cdot x_i \end{cases} \tag{5}$$

$$p_i \cdot v_i = \frac{len(p_{i+1}, p_i)}{p_{i+1} \cdot t_{i+1} - p_i \cdot t_i} \tag{6}$$

where $p_i \cdot dir_i$ represents the driving direction of the trajectory point p_i at time t_i, i.e., the clockwise angle between the straight line of two adjacent trajectory points and the geographic north; $p_i \cdot v_i$ represents the instantaneous speed of point p_i at time t_i; $len(p_i, p_j)$ represents a function for calculating the road network distance between two trajectory points.

Based on the instantaneous speed and direction of the trajectory point, the traffic speed of the road segments $\boldsymbol{L_s} = \{l_i^s\}_{i=1}^{M}$ is extracted. As each road segment has two directions, the dimension of the traffic condition matrix $\boldsymbol{X}$ is $2M * N$. The traffic speed of the $l_i^s \cdot dir_1$ is obtained by $\boldsymbol{X}\left(s_{2i-1}, \{w_k\}_{k=1}^{N}\right)$, and the traffic speed of the $l_i^s \cdot dir_2$ is obtained by $\boldsymbol{X}\left(s_{2i}, \{w_k\}_{k=1}^{N}\right)$. In addition, to alleviate the influence of abnormal values on speed estimation, the harmonic mean (Salamanis et al. 2016), rather than the arithmetic mean, is used to calculate traffic speed of a specific road segment in a time window. For example, the traffic speed on the direction $l_i^s \cdot dir_1$ within the time window $w_k = [t_s, t_e]$ is calculated by Eq. (7).

$$X(s_{2i-1}, w_k) = \frac{n}{\sum_{j=1}^{n} \frac{1}{p_j \cdot v_j}} \; p_j \in l_i^s \cdot dir_1 \cap p_j \cdot t_j \in [t_s, t_e] \tag{7}$$

where $p_j \in l_i^s \cdot dir_1$ indicates that the movement direction of the point is the same as the direction of the road segment and is closer to the road segment l_i^s; t_s represents the start time of the time window w_k; t_e represents the end time of the time window w_k, Note: we only use the vehicle trajectory under the passenger state to calculate the speed of the road segment.

4.3 Sparse traffic speed imputation

Owing to the sparse distribution of trajectory points, the spatiotemporal condition matrix $\boldsymbol{X}$ contain many missing elements. In order to obtain the complete spatiotemporal condition matrix $\widehat{\boldsymbol{X}}$ of the entire road network, the missing values in the spatiotemporal condition matrix $\boldsymbol{X}$ are imputed. Existing studies integrate spatial and temporal correlation models to improve the performance of missing data imputation. However, it is difficult to directly capture the traffic flow patterns from the spatial dimension on the signalized road network. For example, the traffic flow patterns of adjacent segments with different directions are usually different. Therefore, a lightweight multi-view learning method is proposed to improve the performance of missing data imputation by integrating temporal patterns and spatial topological relations. In the time dimension, the imputed results of the

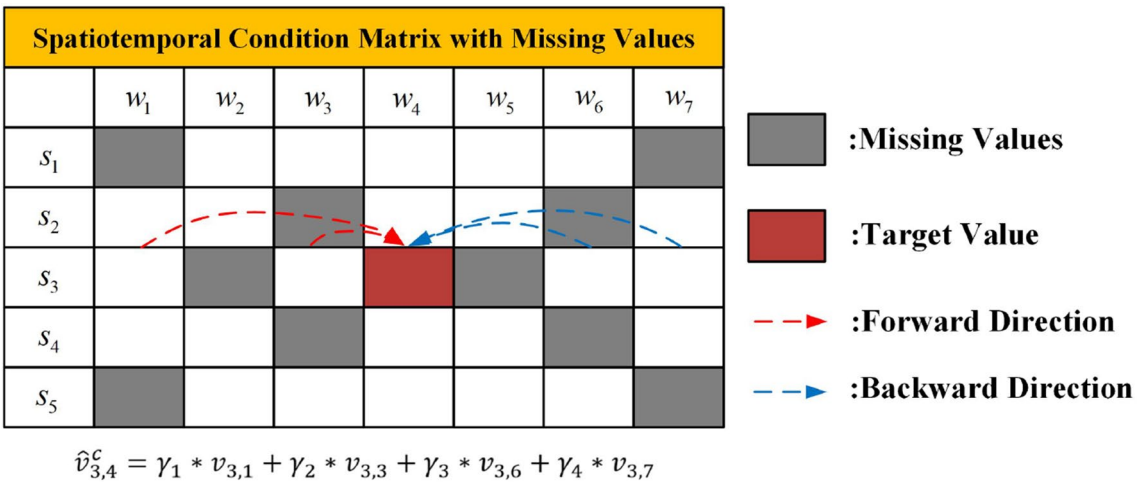

Fig. 6 Illustration of Improved-SES and Improved-MA algorithm

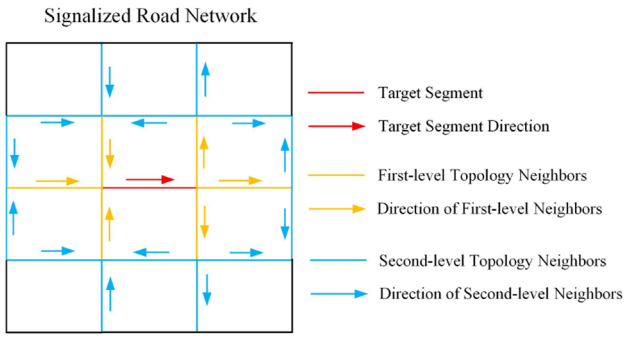

Fig. 7 Topological relations between target segment neighboring segments

three views of closeness, daily periodicity and weekly periodicity are integrated. In the spatial dimension, the imputed results of considering topological connections are integrated.

Prior studies show that the traffic condition has typical closeness, daily periodicity, and weekly periodicity characteristics in the time dimension (Cheng et al. 2019). In the time dimension, sparse data imputation can be transformed into the traditional time series modeling problem, and the missing values can be estimated by using the samples of the adjacent historical times. For the temporal closeness, an improved simple exponential smoothing (Improved-SES) algorithm is used to mine the missing close patterns. For the periodicity, an improved moving average (Improved-MA) algorithm is used to mine the missing periodic patterns. Compared with traditional MA and SES algorithms, Improved-MA and Improved-SES describe the temporal characteristics of traffic flow from two temporal directions (forward and backward).

As shown in Fig. 6, the sample data of forwarding and backward time intervals were selected, using the time interval within which the missing data were located as the center position, which can alleviate the deficiency of the time lag in imputation results of the traditional SES and MA algorithm (Yi et al. 2016). Specifically, if the traffic condition within the time window w_k on the direction $l_i^s \cdot dir_2$ is missing, the imputation values considering the corresponding closeness view is shown in Eq. (8).

$$\begin{cases} \hat{v}^c_{2i,k} = \dfrac{\sum_{j=1}^{\mathcal{L}^c} v_{2i,k-j} * \gamma^c * (1-\gamma^c)^{j-1} + \sum_{j=1}^{\mathcal{L}^c} v_{2i,k+j} * \gamma^c * (1-\gamma^c)^{j-1}}{2\sum_{j=1}^{\mathcal{L}^c} \gamma^c * (1-\gamma^c)^{j-1}} \\ v_{2i,k-j} = \boldsymbol{X}(s_{2i}, w_{k-j}) \\ v_{2i,k+j} = \boldsymbol{X}(s_{2i}, w_{k+j}) \end{cases} \tag{8}$$

where $\hat{v}^c_{2i,k}$ represents the estimated value from the view of closeness; $\mathcal{L}^c$ represents the step of backward and forward dependence; γ^c represents the smoothing parameters, with a value range of [0,1]; Similarly, within the time window w_k on direction $l_i^s \cdot dir_2$, the estimated values from the views of daily periodicity and weekly periodicity are respectively expressed as $\hat{v}^{dp}_{2i,k}$ and $\hat{v}^{wp}_{2i,k}$, and the corresponding dependent steps are respectively expressed as $\mathcal{L}^{dp}$ and $\mathcal{L}^{wp}$.

In the spatial dimension, based on the topological relations $\mathcal{T}_s$, the second-level topological neighbors of the target road segment are identified as the spatial neighboring segments. As shown in Fig. 7, the first-level topological neighbors are directly connected to the target road segment, and the second-level topological neighbors are directly connected to the first-level topological neighbors. It should be noted that topological neighbors not only need to be spatially connected to the target road segment, but also need to be consistent with the filling direction of the target road segment. If the second-level topological neighbors of $l_i^s \cdot dir_2$ are expressed as $\boldsymbol{\Omega}_{\mathcal{T}_s}^{i \to dir_2}$, the estimated result of the traffic condition from the spatial view in time window w_k is shown in Eq. (9).

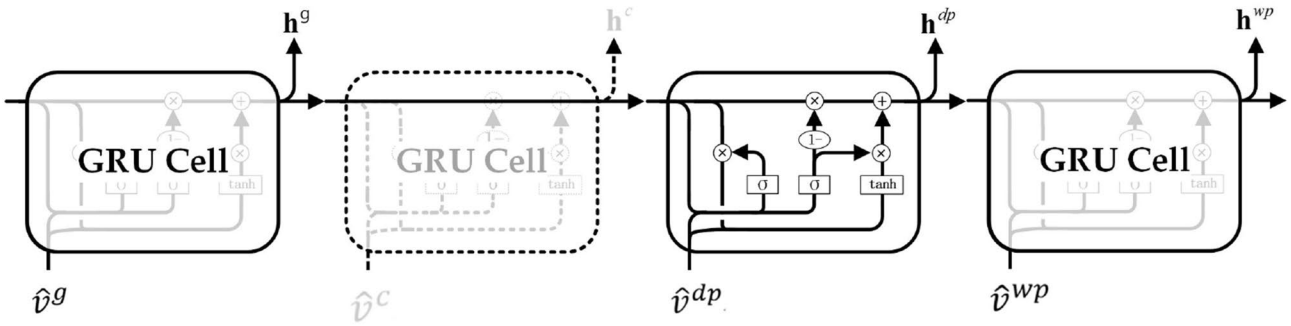

Fig. 8 Fusion of results of multiple views

$$\begin{cases} \hat{v}^{g}_{2i,k} = \dfrac{\sum_{j\in\Omega^{i\to dir_2}_{T_s}} v_{j,k} * \frac{1}{d^{s}_{2i\leftrightarrow j}}}{\sum_{j\in\Omega^{i\to dir_2}_{T_s}} \frac{1}{d^{s}_{2i\leftrightarrow j}}} \\ v_{j,k} = X(s_j, w_k) \\ d^{s}_{2i\leftrightarrow j} = \left| \boldsymbol{X}(s_j) - \boldsymbol{X}(s_{2i}) \right| \end{cases} \tag{9}$$

where $\hat{v}^{g}_{2i,k}$ represents the estimated value from the spatial view; $d^{s}_{2i\leftrightarrow j}$ represents the distance between the sequence $\boldsymbol{X}(s_j)$ and $\boldsymbol{X}(s_{2i})$; $v_{j,k}$ represents the observed values of the object s_j in the time window w_k.

For different views of traffic condition, four different estimation results are obtained, i.e., $\hat{v}^{g}$, $\hat{v}^{c}$, $\hat{v}^{dp}$ and $\hat{v}^{wp}$. Theoretically, the fusion of the four results can improve the accuracy of the final data estimation (Cheng et al. 2020a; b, c; Jiang et al. 2018). However, when the single-view estimation results deviate significantly from the actual values, the simple fusion may worsen estimation accuracy. Therefore, we introduce a filtering mechanism before data fusion, the filtering process of a road segment is shown in inequality (10).

$$\left| v - \hat{v}^{\alpha} \right| < d_{\epsilon} \rightarrow \left| \hat{v}^{\alpha} - \hat{v}^{\beta} \right| < d_{\epsilon} \ \forall \alpha, \beta \in (g, c, dp, wp) \tag{10}$$

where v represents the actual value; $\hat{v}^{g}$ represents the estimation result under the spatial view; $\hat{v}^{c}$ represents the estimation result under the temporal closeness view; $\hat{v}^{dp}$ represents the estimation result under the daily periodicity view; $\hat{v}^{wp}$ represents the estimation result under the weekly periodicity view; $\hat{v}^{\alpha}$ represents the estimated value under a specific view, that is, when the single-view estimation result is good, the deviation between the estimated value and the actual value should be less than d_{ϵ}. Considering that the actual value v cannot be obtained, the constraint $\left| v - \hat{v}^{\alpha} \right| < d_{\epsilon}$ can be simplified as $\left| \hat{v}^{\alpha} - \hat{v}^{\beta} \right| < d_{\epsilon}$, that is, the deviation between any two views is less than d_{ϵ}. The estimated values of the view satisfying the constraints constitute the vector $\hat{\boldsymbol{v}}$, where $1 < |\hat{\boldsymbol{v}}| \leq 4$.

Since $\hat{\boldsymbol{v}}$ has variable lengthes, the fusion method should be able to handle sequences with variable lengths. To this end, we adopt gated recurrent unit network (GRU) (Chung et al. 2014) to fuse the estimation results of the four views. As shown in Fig. 8, if $\hat{\boldsymbol{v}} = \{\hat{v}^{g}, \hat{v}^{dp}, \hat{v}^{wp}\}$, the GRU cell corresponding to $\hat{v}^{c}$ is omitted, and only the estimation results of the three views are fused. The process of single forward propagation is shown in Formula (11).

$$\begin{cases} \boldsymbol{z}^{wp} = \sigma(\boldsymbol{W}_z \hat{\boldsymbol{v}}^{wp} + \boldsymbol{U}_z \boldsymbol{h}^{dp} + \boldsymbol{b}_z) \\ \boldsymbol{r}^{wp} = \sigma(\boldsymbol{W}_r \hat{\boldsymbol{v}}^{wp} + \boldsymbol{U}_r \boldsymbol{h}^{dp} + \boldsymbol{b}_r) \\ \tilde{\boldsymbol{h}}^{dp} = \tanh\left(\boldsymbol{W}_h \hat{\boldsymbol{v}}^{wp} + \boldsymbol{U}_h \left(\boldsymbol{r}^{wp} \odot \boldsymbol{h}^{dp}\right) + \boldsymbol{b}_h\right) \\ \boldsymbol{h}^{wp} = \boldsymbol{z}^{wp} \odot \boldsymbol{h}^{dp} + (1 - \boldsymbol{z}^{wp}) \odot \tilde{\boldsymbol{h}}^{wp} \\ \hat{\boldsymbol{v}} = \sigma(\boldsymbol{W}_o \boldsymbol{h}^{wp} + \boldsymbol{b}_o) \end{cases} \tag{11}$$

where $\boldsymbol{h}^{wp}$ represents the hidden unit; $\boldsymbol{z}^{wp}$ represents the output of the update gate; $\boldsymbol{r}^{wp}$ represents the output of the reset gate; $\hat{\boldsymbol{v}}^{wp}$ represents the estimation result of a single view; $\boldsymbol{W}$, $\boldsymbol{U}$, and $\boldsymbol{b}$ represent the parameters that can be optimized by the model; $\odot$ denotes Hadamard product; $\hat{\boldsymbol{v}}$ represents the output of the model; σ represents the sigmoid activation function.

The network can be trained by minimizing the square loss between the fused traffic speed and the true traffic speed. The loss function is defined in Formula (12).

$$loss = \frac{1}{2} \sum_{(i,j)\in\boldsymbol{\Omega}} \left(\hat{v}_{ij} - v_{ij} \right)^2 \tag{12}$$

where $\boldsymbol{\Omega}$ represents the observable index set; $\hat{v}_{ij}$ represents the result of the fusion; v_{ij} represents the expected output, that is, the actual value.

4.4 Complexity analysis of NEI-SE

In this section, we analyze the time complexity of the NEI-SE. For the proposed approach, missing data imputation is the most time-consuming step. The most time-consuming part of missing data imputation is the forward propagation operation of GRU. Therefore, we mainly analyze the time

Table 1 Samples of taxi trajectory data

Taxi Id	Time	Latitude	Longitude
9634CE[a]	2018-07-01 00:00:01	30.6[a]	114.1[a]
6582DP[a]	2018-07-01 00:00:06	30.5[a]	114.2[a]
1345LE[a]	2018-07-01 00:00:09	30.5[a]	114.2[a]
……	……	……	……
9537EE[a]	2018-08-31 23:59:48	30.5[a]	114.3[a]

[a]Means the content is omitted

complexity of GRU. Suppose the dimension of the learnable weight matrix in GRU is n, the time complexity of one operation of GRU is $O(n^2)$. In this study, the maximum number of iterations of GRU is four. Hence the final time complexity is still $O(n^2)$. At present, the commonly used data imputation method is tensor decomposition method. The tensor decomposition method needs to construct a three-dimensional tensor, and the time complexity of its solution is about equal to $O(n^3)$. Therefore, the proposed approach has obvious computational advantages.

5 Experimental results and discussions

5.1 Data preparation

5.1.1 Data sources

Three kinds of datasets were used to validate the proposed speed estimation approach: (1) traffic network data, (2) taxi trajectory data, and (3) AutoNavi congestion index data. The traffic road network data and taxi trajectory data were used to estimate the traffic speed, and the AutoNavi congestion data were used to evaluate the quality of the estimated traffic speed data.

Collected from the Amap platform, the traffic network data cover main urban roads within the second ring road of Wuhan, China.

Table 2 Samples of AutoNavi congestion data

Congestion id	Time	Shape	Congestion index
b5a6[a]	2018-07-01 00:04:43	Geometry (polyline)	3.6
9259[a]	2018-07-01 00:15:57	Geometry (polyline)	4.4
5d8a[a]	2018-07-01 00:18:35	Geometry (polyline)	10.0
……	……	……	……
5fdd[a]	2018-08-31 23:54:24	Geometry (polyline)	5.7

[a]Means the content is omitted

Table 3 Length distribution characteristics of the road network after segmentation

Length of road segment (m)	Number
$len(l_i^s) < 400$	24
$len(l_i^s) = 400$	405
$400 < len(l_i^s) \le 600$	121
$600 < len(l_i^s) \le 800$	97

The taxi trajectory data were gathered from the Traffic Management Bureau of Wuhan, China. The period of the trajectory data is from July 1, 2018, to August 31, 2018, and the sampling intervals range from 5 to 30 s. The total number of trajectory points in the two months is more than 500 million, and the average number of trajectory points in one day is about 8 million. As shown in Table 1, each trajectory point contains unique identification, recording time, longitude, and latitude. In order to protect privacy, the unique identification of taxi is encrypted.

The AutoNavi congestion data came from the Amap open platform. The Amap API was used to crawl the congestion data of the study region. The period of the congestion data is from July 1, 2018 to August 31, 2018. There are more than 3.7 million congestion events in two months, with an average of 60,000 congestion events in one day. Each AutoNavi congestion data record contains the unique identification and the happening time of a congestion event, the duration of congestion, and the congestion index. The congestion index quantitatively describes the severity of road congestion. Compared with the trajectory data, AutoNavi congestion data have much better quality. Therefore, AutoNavi congestion data can be regarded as the ground truth (Table 2).

5.1.2 Data preprocessing

As road network data, trajectory data, and AutoNavi congestion data were collected from different sources, and we need to preprocess the original data. The preprocessing process consists of two steps:

(1) The coordinate systems of AutoNavi congestion data, road network data, and taxi trajectory data are transformed into the WGS84 coordinate system.
(2) We extracted the congestion data and taxi track data within the second ring road of Wuhan.

5.1.3 Characteristics of extraction speed

When performing road network segmentation, the length threshold d_r was set to 400 m after extensive experiments, and 647 road segments were obtained after segmentation. Table 3 shows the distribution characteristics of road lengths after

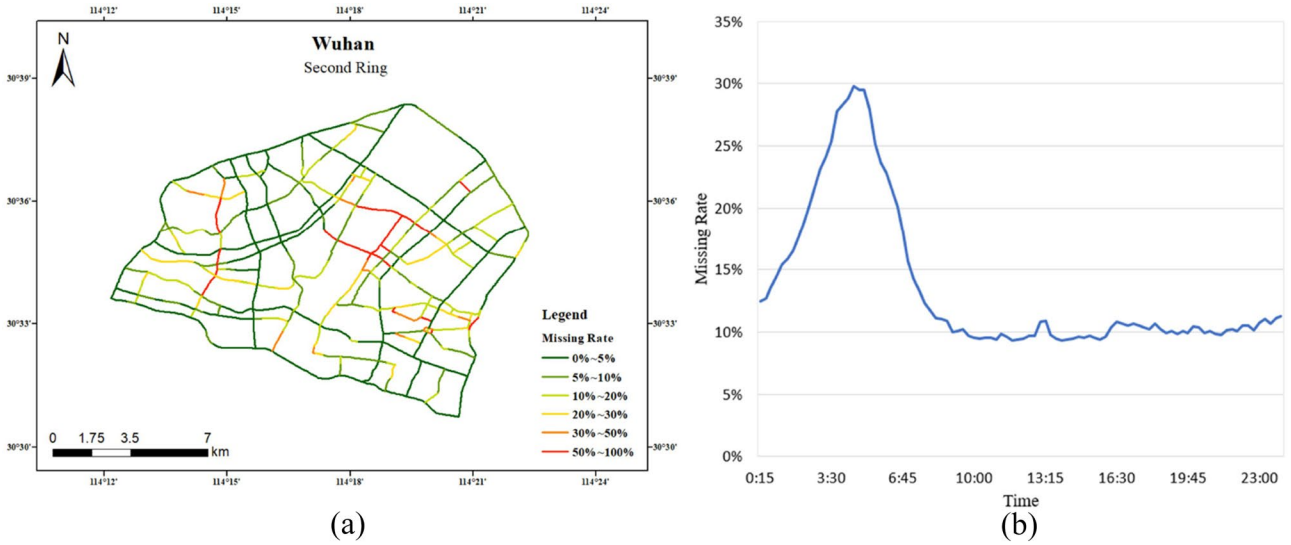

Fig. 9 Missing rate of traffic speed in temporal and spatial dimensions

segmentation. Road segments with a length of less than 400 m account for 62.5% of total road segments, indicating that there are still many long road segments of the road network after segmentation by signalized intersections. To accurately extract the traffic speed of different areas in a long road segment, it is reasonable to divide the road segment further.

In the speed extraction stage, the time window is set to 15 min, similar to Hou et al. (2019). That is, a day contains a total of 96-time windows. After speed extraction, a total of 1,041,839 traffic speed values were obtained for the two months, and the overall missing rate was 13.5%. In addition, we further show the missing rate of speed data from time and space dimensions in Fig. 9. From the spatial dimension, the missing rates of different road segments are heterogeneous. The missing rate of most road segments is less than 15%, and the missing rate of very few road segments is more than 50%. From the perspective of the time dimension, trajectory data within the 15-min window cover 70–90% of the traffic segments at most. This also indicates that sparse data estimation is a crucial step in trajectory-based traffic speed estimation.

5.2 Sparse traffic data recovery

5.2.1 Evaluation metrics of sparse data recovery

In sparse data imputation, a critical problem is how to evaluate the performance of the imputation model. In this study, mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) are used as quantitative indicators to verify the imputation accuracy of the proposed model. The calculation methods of MAE, RMSE and MAPE are shown in Formulas (13), (14), and (15).

$$MAE = \frac{1}{N}\sum_{(i,j)\notin\mathbf{\Omega}}\left|\boldsymbol{X}(s_i,w_j)-\hat{\boldsymbol{X}}(s_i,w_j)\right| \tag{13}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{(i,j)\notin\mathbf{\Omega}}\left(\boldsymbol{X}(s_i,w_j)-\hat{\boldsymbol{X}}(s_i,w_j)\right)^2} \tag{14}$$

$$MAPE = \frac{100\%}{N}\sum_{(i,j)\notin\mathbf{\Omega}}\left|\frac{\boldsymbol{X}(s_i,w_j)-\hat{\boldsymbol{X}}(s_i,w_j)}{\boldsymbol{X}(s_i,w_j)}\right| \tag{15}$$

where $\mathbf{\Omega}$ represents the indexed set of observable data; N represents the total number of missing data; $\boldsymbol{X}(s_i,w_j)$ represents the real traffic speed of the spatial object s_i in the time window w_j; $\hat{\boldsymbol{X}}(s_i,w_j)$ represents the traffic speed estimated by the model within the time window w_j of the spatial object s_i.

5.2.2 Baseline methods

To comprehensively evaluate the performance of proposed approach, we used nine baseline methods for comparison based on two missing types (random missing, block missing) and four missing rates (20, 30, 40, and 50%):

- **HA** (Campbell and Thompson 2008): Historical average (HA) is a statistical model which fills missing traffic condition in each road segment by averaging all the traffic conditions in the historical time slot.
- **SES** (Gardner 2006): Simple exponential smoothing (SES) is a special weighted historical average model

which fills missing traffic condition by weighting the smoothed value and the observed historical value.

- **ST-KNN** (Cai et al. 2016; Yu et al. 2016a, b): Spatiotemporal K-Nearest Neighbors (ST-KNN) is a data-driven model which imputes the missing traffic conditions by searching for k spatiotemporal nearest neighbors in the historical database.
- **ST-ISE** (Cheng et al. 2020a; b, c): Lightweight ensemble spatiotemporal interpolation (ST-ISE) is a data-driven model which imputes missing traffic conditions on each road segment using Extreme Learning Machine to integrate SES and IDW interpolation results.
- **ST-2SMR** (Jiang et al. 2018): Spatiotemporal two-step missing data reconstruction (ST-2SMR) is a data-driven model which uses coarse and fine interpolations to improve the final imputation performance.
- **TRMF** (Yu et al. 2016a, b): Temporal regularized matrix factorization (TRMF) is a data-driven model which incorporates temporal dependencies as a regularization term into matrix factorization to impute missing traffic conditions on each road segment.
- **BTMF** (Chen and Sun 2021): Bayesian temporal matrix factorization (BTMF) is a variation of TRMF model which incorporates Bayesian theory into the solution of TRMF model to impute missing traffic conditions on each road segment.
- **LRTC-TNN** (Chen et al. 2020): Low-rank tensor completion with truncated nuclear norm (LRTC-TNN) is an improved tensor factorization method that imputes the missing traffic conditions by factorizing traffic tensors of location × day × time windows.
- **BTTF** (Chen and Sun 2021): Bayesian temporal tensor factorization (BTTF) is an advanced tensor factorization method that leverages temporal dependencies to impute missing traffic condition based on traditional tensor factorization.

5.2.3 Parameter selection

The hyperparameters in the imputation process of the NEI-SE mainly include the time closeness dependency step $\mathcal{L}^c$, the smoothing parameter γ^c, time daily periodicity dependency step $\mathcal{L}^{dp}$, time weekly periodicity dependency step $\mathcal{L}^{wp}$,

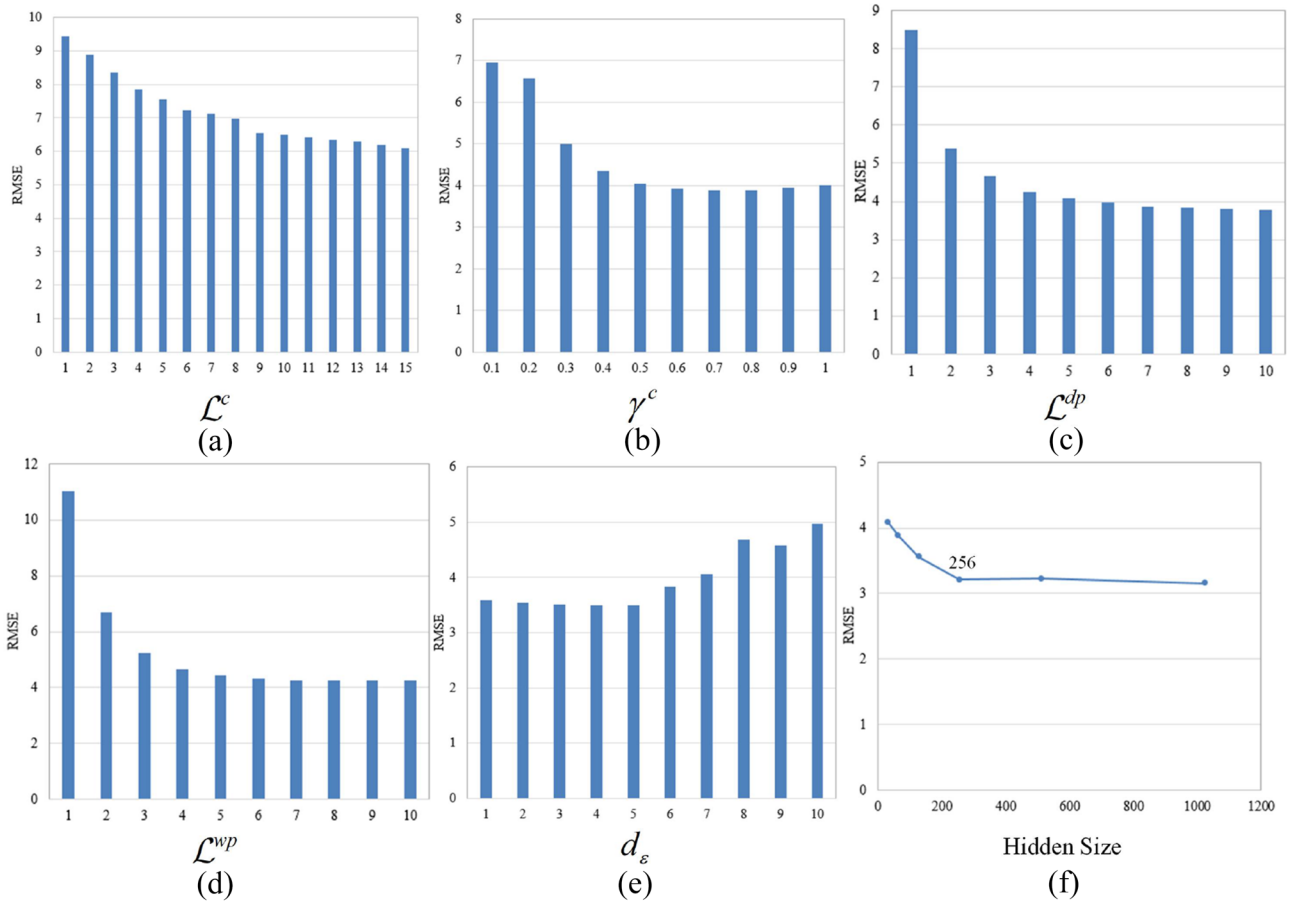

Fig. 10 Parameter tuning of NEI-SE

Table 4 Comparison results (in MAE/RMSE/MAPE) with baselines for random missing

Models	Missing rate			
	20%	30%	40%	50%
HA	3.69/4.96/11.05%	3.78/5.11/11.37%	3.90/5.28/11.77%	4.04/5.48/12.24%
SES	3.28/4.50/9.56%	3.36/4.62/9.80%	3.45/4.77/10.11%	3.57/4.97/10.50%
ST-KNN	2.75/3.64/8.46%	2.81/3.68/8.58%	2.93/3.75/8.72%	3.04/3.84/8.92%
ST-ISE	2.85/3.76/8.59%	2.71/3.55/8.21%	2.77/3.59/8.53%	2.90/3.78/8.89%
ST-2SMR	3.42/4.79/10.20%	3.93/5.47/11.82%	4.06/5.49/12.29%	4.06/6.99/13.05%
TRMF	2.80/3.41/8.47%	2.81/3.42/8.48%	2.81/3.42/8.49%	2.81/3.43/8.50%
BTMF	2.78/3.37/8.28%	2.79/3.38/8.31%	2.80/3.40/8.35%	2.81/3.42/8.39%
LRTC-TNN	2.94/3.70/8.41%	3.02/3.80/8.62%	3.12/3.94/8.88%	3.25/4.11/9.22%
BTTF	2.79/3.36/8.27%	2.79/3.37/8.29%	2.80/3.39/8.33%	2.81/3.41/8.37%
NEI-SE	2.66/3.31/8.14%	2.70/3.55/8.16%	2.69/3.33/8.19%	2.76/3.35/8.24%

Table 5 Comparison results (in MAE/RMSE/MAPE) with baselines for block missing

Models	Missing rate			
	20%	30%	40%	50%
HA	5.84/8.30/17.23%	5.23/8.78/18.61%	6.99/9.87/20.73%	7.38/10.58/21.66%
SES	5.69/8.13/17.35%	6.03/8.46/18.41%	6.63/9.37/20.10%	6.80/9.84/20.51%
ST-KNN	3.13/4.13/9.28%	3.49/4.68/10.49%	3.82/5.25/11.62%	4.60/6.41/14.28%
ST-ISE	3.68/4.89/11.11%	4.56/6.12/14.41%	4.69/6.27/14.75%	5.39/7.20/17.01%
ST-2SMR	4.91/6.59/14.75%	5.14/6.85/15.45%	5.75/7.76/17.31%	6.38/8.25/18.84%
TRMF	3.49/5.08/10.00%	3.74/6.14/10.98%	5.09/9.05/14.04%	5.21/9.35/14.63%
BTMF	3.41/4.85/9.78%	3.77/6.22/11.03%	4.93/8.69/13.65%	5.12/9.21/14.41%
LRTC-TNN	3.26/4.08/9.54%	3.40/4.27/9.87%	3.75/4.69/10.64%	3.90/4.93/11.28%
BTTF	2.88/3.54/8.65%	2.87/3.53/8.70%	2.93/3.62/8.86%	2.96/3.70/8.95%
NEI-SE	2.84/3.64/8.43%	2.90/3.52/8.70%	2.89/3.51/8.68%	2.88/3.54/8.71%

the multi-view filtering threshold d_ϵ, and the number *num* of nodes in GRU hidden layer. In the modeling process, the control variable method is used to obtain the optimal combination of parameters, in which the value range of γ^c is [0.1, 1], the value range of $\mathcal{L}^c$ is [1, 15], the value range of num is [32, 62, 128, 256]. The value range of $\mathcal{L}^{dp}$, $\mathcal{L}^{wp}$, and d_ϵ is [1, 15]. Taking the random missing rate of 30% as an example, Fig. 10 shows the process of parameter calibration. The results show that with the increase of $\mathcal{L}^c$ and γ^c, the imputation accuracy of the NEI-SE gradually increases. When $\mathcal{L}^c = 9$ and $\gamma^c = 0.5$, the model achieves good accuracy. Similarly, $\mathcal{L}^{dp}$, $\mathcal{L}^{wp}$, d_ϵ, and *num* are also calibrated. Finally, the following settings of parameters were used throughout the experiments: $\mathcal{L}^{dp} = 5$, $\mathcal{L}^{wp} = 4$, $d_\epsilon = 5$, $num = 256$.

5.2.4 Comparison with baselines

In this section, we compare the imputation performance of the NEI-SE with baselines. Table 4 shows the comparison results between NEI-SE and baselines under random missing. The results show that the imputation performance of all models decreases slightly with the increase of the missing rate. Among them, the imputation performance of the ST-2SMR model is greatly affected by the missing rate. Except for the ST-2SMR model, other models were less affected by the missing rate. In addition, NEI-SE achieves the best imputation performance under random missing. Table 5 shows the comparison results between NEI-SE and baselines under block missing. The results show that with the increase of missing rate, the imputation performance of most models (such as ST-KNN, ST-ISE, ST-2SMR, TRMF, and BTMF) decreases significantly, while the imputation performance of the tensor decomposition model and NEI-SE is relatively stable. Compared with BTTF, NEI-SE has less physical storage space. The reason is that BTTF is a tensor decomposition model, and each solution needs to construct a tensor of location × day × time windows.

5.2.5 Imputation performance of the NEI-SE

In this section, the scatter plots are used to describe the performance of NEI-SE qualitatively. Figure 11 shows the imputation results of the NEI-SE with a single road segment with a missing rate of 30–40%. The results show that the observed values are close to the estimated values of the NEI-SE, and the

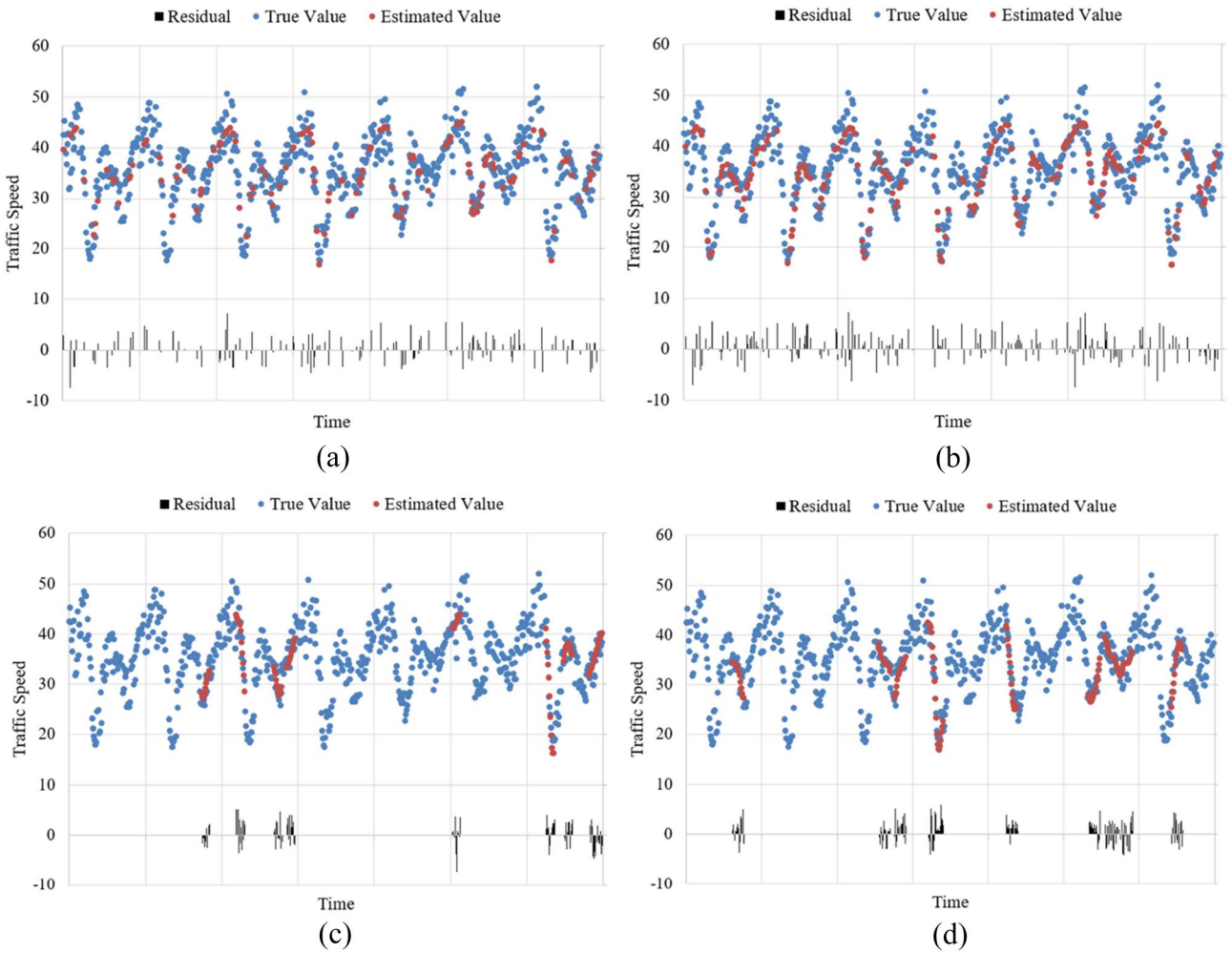

Fig. 11 Estimated values and corresponding actual values of a road segment: **a** 30% random missing, **b** 40% random missing, **c** 30% block missing, and **b** 40% block missing

Table 6 Imputation comparison (in MAE/RMSE/MAPE) of different fusion methods under random missing

Models	Missing rate			
	20%	30%	40%	50%
MLP-Fusion	3.93/5.15/11.67%	4.01/5.25/11.69%	4.05/5.31/11.94%	4.33/5.62/12.78%
ELM-Fusion	4.83/6.45/13.76%	4.94/6.59/13.96%	5.06/6.75/14.26%	5.17/6.88/14.53%
Mean-Fusion	2.80/3.48/8.41%	2.82/3.51/8.43%	2.84/3.56/8.50%	2.87/3.61/8.58%
NEI-SE	2.66/3.31/8.14%	2.70/3.55/8.16%	2.69/3.33/8.19%	2.76/3.35/8.24%

Table 7 Imputation comparison (in MAE/RMSE/MAPE) of different fusion methods under block missing

Models	Missing rate			
	20%	30%	40%	50%
MLP-Fusion	4.16/5.41/12.42%	4.92/6.30/14.77%	5.17/6.61/15.25%	5.69/7.21/17.02%
ELM-Fusion	3.95/5.27/11.46%	4.71/6.30/13.38%	5.41/7.19/15.17%	6.46/8.39/18.17%
Mean-Fusion	2.85/3.65/8.59%	2.90/3.81/8.86%	2.92/3.93/8.82%	3.09/4.58/9.02%
NEI-SE	2.84/3.64/8.43%	2.90/3.52/8.70%	2.89/3.51/8.68%	2.88/3.54/8.71%

residuals are mainly between [−5, 5], which indicates that the imputation results of the NEI-SE have high accuracy.

5.2.6 Effect of fusion methods on imputation performance

In the data fusion stage, GRU was used to fuse multi-view estimation results. Therefore, we compare the effects of different fusion methods on imputation performance. We mainly compare four different fusion methods, including the fusion method of multi-layer perceptrons (MLP-Fusion) (Jiang et al. 2018), the fusion method of extreme learning machines (ELM-Fusion) (Cheng et al. 2020a, b, c), average fusion method after filtering (Mean-Fusion), and GRU fusion method after filtering (NEI-SE). Table 6 describes the imputation results of different fusion methods in the scenario of random missing. The results show that NEI-SE has the best imputation accuracy, while ELM-Fusion has the worst imputation performance. At the same time, it can be noted that the imputation accuracy of Mean-Fusion is higher than those of MLP-Fusion and ELM-Fusion, proving the importance of multi-view alignment. In addition, the imputation accuracy of NEI-SE is slightly higher than that of Mean-Fusion, showing that GRU is well suited for data fusion. Similarly, Table 7 describes the imputation results of different fusion methods in the scenario of block missing. The results show that NEI-SE also has the highest imputation accuracy under block-missing scenarios.

5.3 Data quality assessment based on AutoNavi congestion data

5.3.1 Evaluation metric of data quality

We used Pearson correlation coefficient to evaluate the quality of estimated traffic speed data. Pearson correlation coefficient is used to calculate the correlation between AutoNavi congestion index and estiamted traffic speed. The traffic condition matrix of AutoNavi congestion index is represented

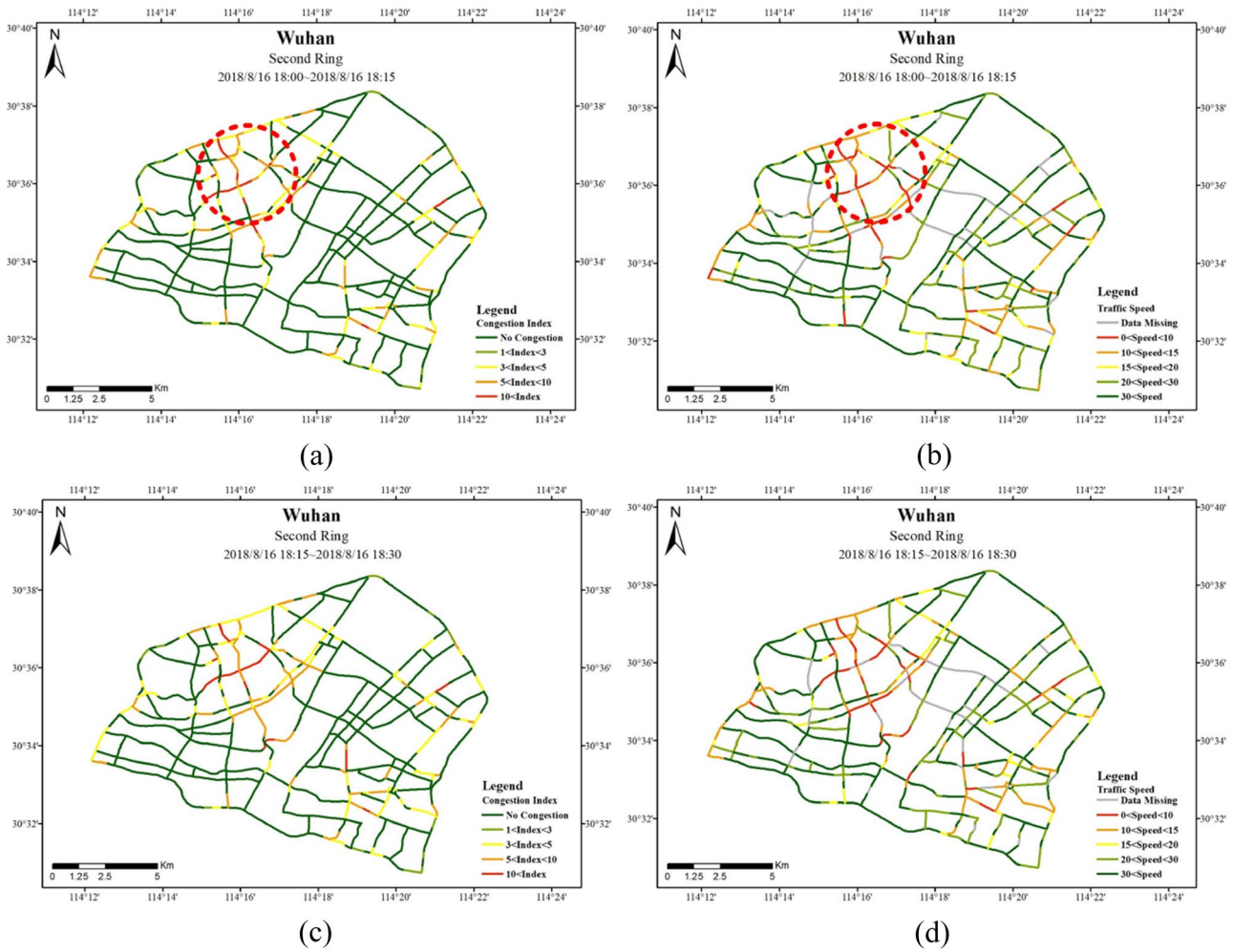

Fig. 12 Spatial distribution of congestion index and traffic speed

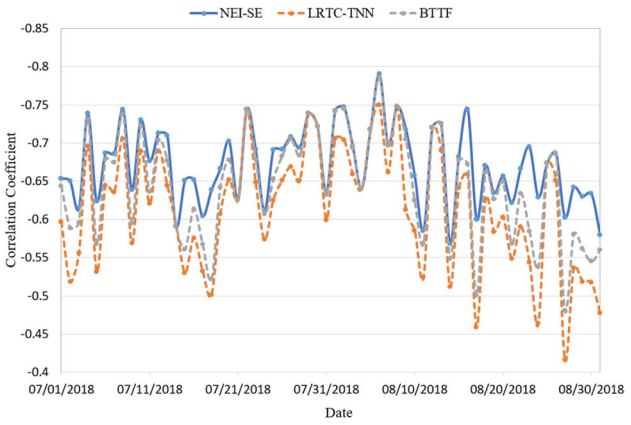

Fig. 13 Changes in correlation coefficient between congestion index and traffic speed over time

by $\boldsymbol{X}_g$, and the traffic condition matrix of traffic speed is represented by $\boldsymbol{X}_v$. The calculation method of the correlation coefficients of $\boldsymbol{X}_v$ and $\boldsymbol{X}_g$ is shown in Formula (16).

$$\begin{cases} r = \dfrac{\sum_{(i,j)\in\boldsymbol{\Omega}} \left(X_v(s_i,w_j)-\overline{X_v(\boldsymbol{\Omega})}\right)\left(X_g(s_i,w_j)-\overline{X_g(\boldsymbol{\Omega})}\right)}{\sqrt{\sum_{(i,j)\in\boldsymbol{\Omega}} \left(X_v(s_i,w_j)-\overline{X_v(\boldsymbol{\Omega})}\right)^2}\sqrt{\sum_{(i,j)\in\boldsymbol{\Omega}} \left(X_g(s_i,w_j)-\overline{X_g(\boldsymbol{\Omega})}\right)^2}} \\ \overline{\boldsymbol{X}_v(\boldsymbol{\Omega})} = \frac{1}{T}\sum_{(i,j)\in\boldsymbol{\Omega}} \boldsymbol{X}_v\left(s_i,w_j\right) \\ \overline{\boldsymbol{X}_g(\boldsymbol{\Omega})} = \frac{1}{T}\sum_{(i,j)\in\boldsymbol{\Omega}} \boldsymbol{X}_g\left(s_i,w_j\right) \end{cases} \tag{16}$$

where $\boldsymbol{\Omega}$ represents an index set that can observe both the congestion index and the speed; T represents the number of observable data, i.e., $T = |\boldsymbol{\Omega}|$; $\boldsymbol{X}_v\left(s_i,w_j\right)$ represents the traffic speed of the spatial object s_i in the time window w_j; $\boldsymbol{X}_g\left(s_i,w_j\right)$ represents the traffic congestion index of the spatial object s_i in the time window w_j; r represents the correlation coefficient between traffic congestion index and traffic speed, and its value range is $[-1, 1]$; In the real world, the traffic congestion index and traffic speed should have an obvious negative correlation, i.e., $r < 0$. In this study, the smaller the r value, the higher the data quality of traffic speed.

5.3.2 Analysis of data quality assessment results

Figure 12 shows the spatial distribution of the AutoNavi congestion index and traffic speed in two-time windows. The results show that the AutoNavi congestion index and the traffic speed of the two-time windows show a high degree of consistency in the spatial distribution. In other words, the traffic speed in the spatial area with a higher congestion index is lower. In comparison, the traffic speed in the spatial area with a smaller congestion index is higher. For example, the AutoNavi congestion indices of most road segments in the area marked with the red circle in Fig. 12a are more significant than 3. The traffic speeds of most corresponding road segments in Fig. 12b are less than 15 km/h, proving that the approach proposed in this study is practical for extracting traffic speed data based on trajectories.

From the results of sparse trajectory imputation, it can be seen that the three methods of BTTF, LRTC-TNN, and NEI-SE have good results. Therefore, we used the Pearson correlation coefficient to evaluate the quality of traffic speed data quantitatively, and the results are shown in Fig. 13. The results show that the AutoNavi congestion index has a high negative correlation with estimated traffic speed. The average correlation coefficient of the NEI-SE imputation dataset is −0.67, the average correlation coefficient of the BTTF imputation dataset is −0.65, and the average correlation coefficient of the LRTC-TNN imputation dataset is −0.61. There are two main reasons why the correlation coefficient between congestion index and traffic speed is not less than −0.8. First, the function mapping between AutoNavi congestion index and traffic speed may not be linear. Second, there are some errors in the traffic speed extracted based on trajectory. Even so, it still shows that the traffic speed data extracted has high data quality.

6 Conclusions and future work

First, the urban traffic network is divided according to the signalized intersection and 400 m distance, considering the influence of signal timing on urban traffic speed. Then, the traffic condition matrix is extracted based on the taxi trajectory and the divided road network. Finally, a lightweight multi-view learning method integrating temporal patterns and spatial topological relations is proposed to fill the missing values of the traffic condition matrix. We used nine baseline methods for comparison based on two missing types and four missing rates. The results showed that NEI-SE outperformed the nine existing baselines regarding imputation accuracy.

The AutoNavi Congestion data was used to evaluate the data quality of estimated traffic speed data. The results show that the spatial distribution of congestion index and traffic speed is consistent, that is, the traffic speed in the spatial area with higher congestion index is lower, while the traffic speed in the spatial area with lower congestion index is higher. In addition, taking daily as the time interval, we used Pearson correlation coefficient to analyze the correlation between congestion index and estimated traffic speed. The results show that the congestion index has a significant negative correlation with estimated traffic speed, with an average correlation coefficient of −0.67. This not only indicates that the traffic speed data extracted has high data quality, but also proves the effectiveness of the approach proposed.

The proposed approach provides a solution for the traffic condition estimation of the urban signalized road

networks. The proposed approach has three main advantages. First, through the proposed approach, the traffic condition can be estimated through massive vehicle trajectories without installing a large area of sensor equipment. Second, the data imputation method in NEI-SE can be regarded as an independent component and can be applied to different missing data imputation tasks. Three, the proposed approach is easy to implement, and we provide an open-source code implementation.

Although the proposed method has many advantages, it also has limitations. First, we did not integrate the signal timing data into the proposed method due to the lack of signal timing data. Second, the proposed approach is a hybrid model rather than an end-to-end model, which affects the performance of the proposed approach to some extent. Finally, we used GRU for data fusion and the fusion method can be further explored. Given the above problems, future work will focus on two aspects. First, multi-source data will be collected to improve the effectiveness of the approach. Then, a more appropriate data fusion method will be designed to improve the performance of the approach.

Funding This project was supported by National Key R&D Program of China (International Scientific & Technological Cooperation Program) under Grant 2019YFE0106500, National Natural Science Foundation of China under Grant 41871308.

Data and codes availability statement The data and codes that support the findings of this study are available in 'figshare.com' with the identifier https://doi.org/10.6084/m9.figshare.14946009.

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

Angayarkanni SA, Sivakumar R, Ramana Rao YV (2021) Hybrid Grey Wolf: Bald Eagle search optimized support vector regression for traffic flow forecasting. J Ambient Intell Humaniz Comput 12(1):1293–1304. https://doi.org/10.1007/s12652-020-02182-w

Aryaputera AW, Yang D, Zhao L, Walsh WM (2015) Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging. Sol Energy 122:1266–1278. https://doi.org/10.1016/j.solener.2015.10.023

Cai P, Wang Y, Lu G, Chen P, Ding C, Sun J (2016) A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. Transp Res Part C Emerg Technol 62:21–34. https://doi.org/10.1016/j.trc.2015.11.002

Campbell JY, Thompson SB (2008) Predicting excess stock returns out of sample: can anything beat the historical average? Rev Financ Stud 21(4):1509–1531. https://doi.org/10.1093/rfs/hhm055

Chen X, Sun L (2021) Bayesian temporal factorization for multidimensional time series prediction. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2021.3066551

Chen X, He Z, Wang J (2018) Spatial-temporal traffic speed patterns discovery and incomplete data recovery via SVD-combined tensor decomposition. Transp Res Part C Emerg Technol 86:59–77. https://doi.org/10.1016/j.trc.2017.10.023

Chen X, Yang J, Sun L (2020) A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation. Transp Res Part C Emerg Technol 117:102673. https://doi.org/10.1016/j.trc.2020.102673

Cheng S, Lu F (2017) A two-step method for missing spatio-temporal data reconstruction. ISPRS Int J Geo Inf 6(7):187. https://doi.org/10.3390/ijgi6070187

Cheng S, Lu F, Peng P, Wu S (2019) Multi-task and multi-view learning based on particle swarm optimization for short-term traffic forecasting. Knowl-Based Syst 180:116–132. https://doi.org/10.1016/j.knosys.2019.05.023

Cheng S, Lu F, Peng P (2020a) Short-term traffic forecasting by mining the non-stationarity of spatiotemporal patterns. IEEE Trans Intell Transp Syst. https://doi.org/10.1109/TITS.2020.2991781

Cheng S, Peng P, Lu F (2020b) A lightweight ensemble spatiotemporal interpolation model for geospatial data. Int J Geogr Inf Sci 34(9):1849–1872. https://doi.org/10.1080/13658816.2020.1725016

Cheng S, Zhang B, Peng P, Yang Z, Lu F (2020c) Spatiotemporal evolution pattern detection for heavy-duty diesel truck emissions using trajectory mining: a case study of Tianjin, China. J Clean Prod 244:118654. https://doi.org/10.1016/j.jclepro.2019.118654

Cheng S, Lu F, Peng P, Zheng J (2021) Emission characteristics and control scenario analysis of VOCs from heavy-duty diesel trucks. J Environ Manag 293:112915. https://doi.org/10.1016/j.jenvman.2021.112915

Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. http://arxiv.org/abs/1412.3555 [Cs]

Gardner ES (2006) Exponential smoothing: the state of the art—Part II. Int J Forecast 22(4):637–666. https://doi.org/10.1016/j.ijforecast.2006.03.005

González CL, Zapotecatl JL, Gershenson C, Alberola JM, Julian V (2020) A robustness approach to the distributed management of traffic intersections. J Ambient Intell Humaniz Comput 11(11):4501–4512. https://doi.org/10.1007/s12652-019-01424-w

Guo Q, Li L, Ban X (2019a) Urban traffic signal control with connected and automated vehicles: a survey. Transp Res Part C Emerg Technol 101:313–334. https://doi.org/10.1016/j.trc.2019.01.026

Guo S, Lin Y, Li S, Chen Z, Wan H (2019b) Deep spatial-temporal 3D convolutional neural networks for traffic data forecasting. IEEE Trans Intell Transp Syst 20(10):14

Hara Y, Suzuki J, Kuwahara M (2018) Network-wide traffic state estimation using a mixture Gaussian graphical model and graphical lasso. Transp Res Part C Emerg Technol 86:622–638. https://doi.org/10.1016/j.trc.2017.12.007

Hou Q, Leng J, Ma G, Liu W, Cheng Y (2019) An adaptive hybrid model for short-term urban traffic flow prediction. Physica A 527:121065. https://doi.org/10.1016/j.physa.2019.121065

Hu M-G, Wang J-F, Zhao Y, Jia L (2013) A B-SHADE based best linear unbiased estimation tool for biased samples. Environ Model Softw 48:93–97. https://doi.org/10.1016/j.envsoft.2013.06.011

Jiang L, Zhang X, Zuo W, Xu H, Zhao J, Qiu X, Tian Y, Zhu Y (2018) A neural network method for the reconstruction of winter wheat yield series based on spatio-temporal heterogeneity. Comput Electron Agric 154:46–53. https://doi.org/10.1016/j.compag.2018.08.047

Li L, Losser T, Yorke C, Piltner R (2014) Fast inverse distance weighting-based spatiotemporal interpolation: a web-based application of interpolating daily fine particulate matter PM2:5 in the contiguous US using parallel programming and k-d tree. Int J Environ Res Public Health 11(9):9101–9141. https://doi.org/10.3390/ijerph110909101

Li L, Du B, Wang Y, Qin L, Tan H (2020) Estimation of missing values in heterogeneous traffic data: application of multimodal deep learning model. Knowl-Based Syst 194:105592. https://doi.org/10.1016/j.knosys.2020.105592

Ma D, Luo X, Li W, Jin S, Guo W, Wang D (2017) Traffic demand estimation for lane groups at signal-controlled intersections using travel times from video-imaging detectors. IET Intel Transp Syst 11(4):222–229. https://doi.org/10.1049/iet-its.2016.0233

Praveen DS, Raj DP (2021) Smart traffic management system in metropolitan cities. J Ambient Intell Humaniz Comput 12(7):7529–7541. https://doi.org/10.1007/s12652-020-02453-6

Salamanis A, Kehagias DD, Filelis-Papadopoulos CK, Tzovaras D, Gravvanis GA (2016) Managing spatial graph dependencies in large volumes of traffic data for travel-time prediction. IEEE Trans Intell Transp Syst 17(6):1678–1687. https://doi.org/10.1109/TITS.2015.2488593

Shang J, Zheng Y, Tong W, Chang E, Yu Y (2014) Inferring gas consumption and pollution emission of vehicles throughout a city. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1027–1036. https://doi.org/10.1145/2623330.2623653

Tang J, Zhang X, Yin W, Zou Y, Wang Y (2020a) Missing data imputation for traffic flow based on combination of fuzzy neural network and rough set theory. J Intell Transp Syst. https://doi.org/10.1080/15472450.2020.1713772

Tang K, Tan C, Cao Y, Yao J, Sun J (2020b) A tensor decomposition method for cycle-based traffic volume estimation using sampled vehicle trajectories. Transp Res Part C Emerg Technol 118:102739. https://doi.org/10.1016/j.trc.2020.102739

Tao S, Manolopoulos V, Rodriguez S, Rusu A (2012) Real-time urban traffic state estimation with A-GPS mobile phones as probes. J Transp Technol 2(1):22–31. https://doi.org/10.4236/jtts.2012.21003

Vigos G, Papageorgiou M (2010) A simplified estimation scheme for the number of vehicles in signalized links. IEEE Trans Intell Transp Syst 11(2):312–321. https://doi.org/10.1109/TITS.2010.2042807

Wang Y, Zheng Y, Xue Y (2014) Travel time estimation of a path using sparse trajectories. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp 25–34. https://doi.org/10.1145/2623330.2623656

Wang P, Hu T, Gao F, Wu R, Guo W, Zhu X (2022a) A hybrid data-driven framework for spatiotemporal traffic flow data imputation. IEEE Internet Things J. https://doi.org/10.1109/JIOT.2022.3151238

Wang P, Zhang T, Zheng Y, Hu T (2022b) A multi-view bidirectional spatiotemporal graph network for urban traffic flow imputation. Int J Geogr Inf Sci. https://doi.org/10.1080/13658816.2022.2032081

Wilby MR, Díaz JJV, Gonźlez ABR, Sotelo MÁ (2014) Lightweight occupancy estimation on freeways using extended floating car data. J Intell Transp Syst 18(2):149–163. https://doi.org/10.1080/15472450.2013.801711

Xie P, Li T, Liu J, Du S, Yang X, Zhang J (2020) Urban flow prediction from spatiotemporal data using machine learning: a survey. Inf Fusion 59:1–12. https://doi.org/10.1016/j.inffus.2020.01.002

Xu C-D, Wang J-F, Hu M-G, Li Q-X (2013) Interpolation of missing temperature data at meteorological stations using P-BSHADE*. J Clim 26(19):7452–7463. https://doi.org/10.1175/JCLI-D-12-00633.1

Xu D, Wei C, Peng P, Xuan Q, Guo H (2020) GE-GAN: a novel deep learning framework for road traffic state estimation. Transp Res Part C Emerg Technol 117:102635. https://doi.org/10.1016/j.trc.2020.102635

Yang B, Kang Y, Yuan Y, Huang X, Li H (2021) ST-LBAGAN: spatio-temporal learnable bidirectional attention generative adversarial networks for missing traffic data imputation. Knowl-Based Syst 215:106705. https://doi.org/10.1016/j.knosys.2020.106705

Yi X, Zheng Y, Zhang J, Li T (2016) ST-MVL: filling missing values in geo-sensory time series data. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence, pp 2704–2710

Younes MB (2021) Real-time traffic distribution prediction protocol (TDPP) for vehicular networks. J Ambient Intell Humaniz Comput 12(8):8507–8518. https://doi.org/10.1007/s12652-020-02585-9

Yozgatligil C, Aslan S, Iyigun C, Batmaz I (2013) Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theor Appl Climatol 112(1):143–167. https://doi.org/10.1007/s00704-012-0723-x

Yu B, Song X, Guan F, Yang Z, Yao B (2016a) K-nearest neighbor model for multiple-time-step prediction of short-term traffic condition. J Transp Eng 142(6):04016018. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000816

Yu H-F, Rao N, Dhillon IS (2016b) Temporal regularized matrix factorization for high-dimensional time series prediction. In: 30th conference on neural information processing systems (NIPS 2016b), p 15

Yu B, Yin H, Zhu Z (2018) Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence, pp 3634–3640. https://doi.org/10.24963/ijcai.2018/505

Yu J, Stettler MEJ, Angeloudis P, Hu S, Chen X (2020) Urban network-wide traffic speed estimation with massive ride-sourcing GPS traces. Transp Res Part C Emerg Technol 112:136–152. https://doi.org/10.1016/j.trc.2020.01.023

Zhan X, Li R, Ukkusuri SV (2015) Lane-based real-time queue length estimation using license plate recognition data. Transp Res Part C Emerg Technol 57:85–102. https://doi.org/10.1016/j.trc.2015.06.001

Zhan X, Zheng Y, Yi X, Ukkusuri SV (2017) Citywide traffic volume estimation using trajectory data. IEEE Trans Knowl Data Eng 29(2):272–285. https://doi.org/10.1109/TKDE.2016.2621104

Zhang Z, Li M, Lin X, Wang Y (2020) Network-wide traffic flow estimation with insufficient volume detection and crowdsourcing data. Transp Res Part C Emerg Technol 121:102870. https://doi.org/10.1016/j.trc.2020.102870

Zhao Y, Zheng J, Wong W, Wang X, Meng Y, Liu HX (2019) Various methods for queue length and traffic volume estimation using probe vehicle trajectories. Transp Res Part C Emerg Technol 107:70–91. https://doi.org/10.1016/j.trc.2019.07.008